View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Yandy Scientific Press

Advances in Geo-Energy Research

Vol. 4, No. 2, p. 135-151, 2020

Original article

High-precision calculation of gas saturation in organic shale pores using an intelligent fusion algorithm and a multi-mineral model

Linqi Zhu[1,2], Chaomo Zhang[1,2]*, Zhansong Zhang[1,2], Xueqing Zhou[1,2]

[1] *Key Laboratory of Exploration Technologies for Oil and Gas Resources, Yangtze University, Wuhan 430100, P. R. China*

[2] *Hubei Cooperative Innovation Center of Unconventional Oil and Gas, Yangtze University, Wuhan 430100, P. R. China*

Keywords:
Shale gas
gas saturation
model-driven
data-driven
intelligent fusion algorithm

Cited as:
Zhu, L., Zhang, C., Zhang, Z., Zhou, X. High-precision calculation of gas saturation in organic shale pores using an intelligent fusion algorithm and a multi-mineral model. *Advances in Geo-Energy Research*, 2020, 4(2): 135-151, doi: 10.26804/ager.2020.02.03.

Abstract:
Shale gas reservoirs have been the subject of intensifying research in recent years. In particular, gas saturation has received considerable attention as a key parameter reflecting the gas-bearing properties of reservoirs. However, no mature model exists for calculating the saturation of shale gas reservoirs due to the difficulty in calculating the gas saturation. This paper proposes a new gas saturation prediction method that combines model-driven and data-driven approaches. A multi-mineral petrophysical model is applied to derive the apparent saturation model. Using the calculated apparent saturation, matrix parameters and porosity curve as inputs, an intelligent fusion algorithm composed of five regression algorithms is employed to predict the gas saturation. The gas saturation prediction results in the Yongchuan block, Sichuan Basin, reveal that the model proposed in this paper boasts good reliability and a greatly improved prediction accuracy. The proposed model can greatly assist in calculating the gas saturation of shale gas reservoirs.

1. Introduction

With their continuous exploration and development, shale gas reservoirs have been the subject of increasing attention (Dong et al., 2016; He et al., 2017; Kim et al., 2019; Owusu et al., 2019), particularly as they currently represent fields of new reserves, especially for countries such as China that already host large conventional reservoirs. Accordingly, many countries, including China, the United States, Mexico, Argentina, South Africa, Australia, Canada, Poland, and France, have begun to explore and develop shale gas reservoirs (Yin et al., 2016; Morga and Kamińska, 2018; Jiang et al., 2019a), and hence, research on shale gas reservoirs is incredibly extensive.

The gas-bearing properties of shale gas reservoirs are key parameters for evaluating the quality of those reservoirs (Sang et al., 2018; Tathed and Misra, 2018; Li et al., 2019). Among those parameters, gas saturation is especially indispensable for assessing reservoir quality and calculating other parameters, such as the free gas content, adsorbed gas content and reserves (Ji et al., 2017; Josh, 2019; Lai et al., 2019). In the existing method used to accurately evaluate the gas saturation within a shale gas reservoir at a certain depth, a core is directly extracted over a depth interval, and the obtained core is subjected to an experiment to determine the value of the gas saturation. However, coring and testing are extremely time-consuming and laborious tasks; consequently, it is impossible to core and test each new well. Therefore, core data are typically applied to establish relationships with log responses, and log curves are then utilized to calculate the continuous gas saturation of the entire reservoir.

At present, the gas saturation is calculated mainly by using the resistivity curve and evaluated by utilizing the rock resistivity characteristics, although the non-resistivity curve and the theoretical natural gas response characteristics

*Corresponding author.
E-mail address: 445364694@qq.com (L. Zhu); Zhang7801@263.net (C. Zhang); zhangzhs@yangtzeu.edu.cn (Z. Zhang); 510189613@qq.com (X. Zhou).

Ausasia Science and Technology Press

2207-9963 © The Author(s) 2020.

Received March 12, 2020; revised March 18, 2020; accepted March 24, 2020; available online March 29, 2020.

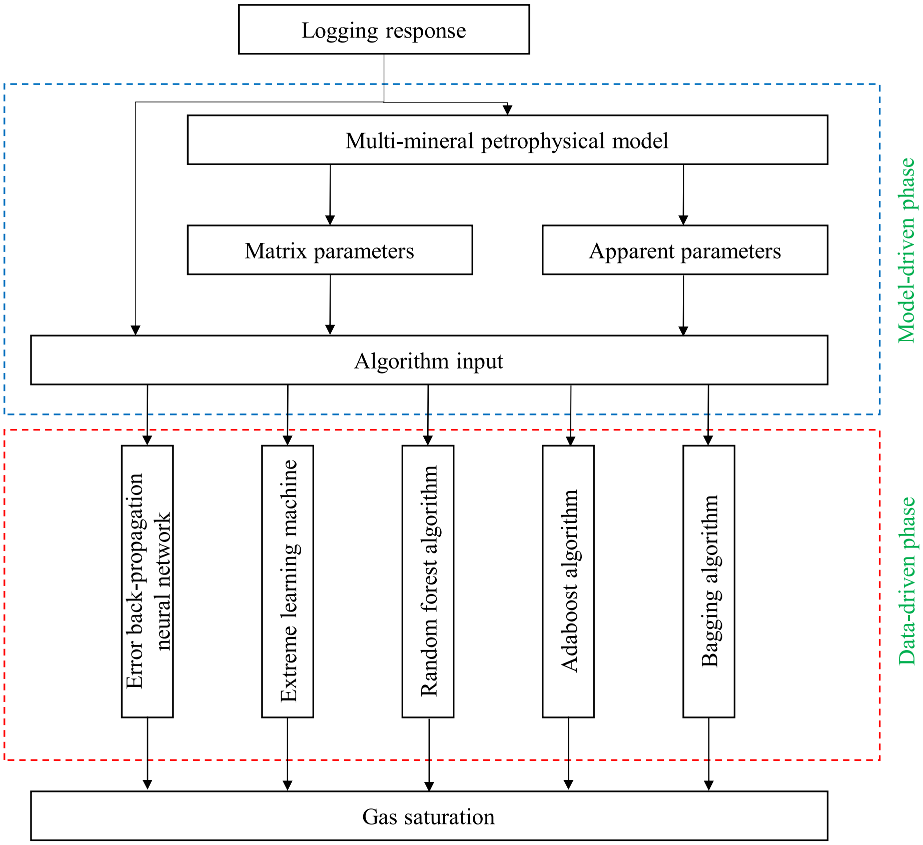


Fig. 1. Gas saturation prediction process.

are also employed (Cai et al., 2017). Eventually, these data are compared with a small amount of core data. For the saturation evaluation method based on the resistivity curve and rock resistivity characteristics, the Archie model is the most mainstream approach (Li et al., 2017; Nazemi et al., 2018; Malekimostaghi et al., 2019; Nazemi et al., 2019). At present, the Archie model is extensively used to evaluate the gas saturation. Various other saturation models, such as the Indonesia model (Chai et al., 2010; Zhou et al., 2019), the Simandoux model (Mashaba et al., 2015; Shedid et al., 2017), the dual-water model (Li et al., 2012; Wu et al., 2019; Tariq et al., 2020), and the gulf effect model (Zhang et al., 2009), are available as alternatives. However, research has indicated that the accuracy of the shale gas reservoir saturation is not high even if the current saturation model based on the conductivity principle is improved. Because the conductivity of a shale gas reservoir is quite complex, research on this topic is incomplete. In response, some scholars have proposed saturation models based on non-resistivity characteristics, such as the density calculation model based on density characteristics proposed by Alfred et al. (2013) and Zhu et al. (2019a), and the uranium content model proposed by Liu et al. (2017). Another technique is the saturation calculation model; however, few studies have applied the saturation calculation model to shale gas reservoirs, and the accuracy of the saturation calculations cannot be guaranteed (Guo et al., 2019). Moreover, since the saturation parameter is more microscopic than the parameters, such as porosity, an algorithm-based version of the saturation evaluation model has not been reported.

In view of the above problems, this paper proposes a gas saturation calculation technique based on an intelligent fusion algorithm and a multi-mineral model. First, using a shale multi-mineral model, two new equations for calculating the apparent gas saturation without calculating the porosity are derived. Then, we propose an intelligent fusion algorithm that combines multiple algorithms to improve the accuracy and emphasize that conventional logging curves, the apparent saturation (the theoretical saturation calculated using petrophysical models), mineral parameters and porosity curves are taken as the inputs for the model.

2. Methodology

The saturation prediction model proposed in this paper consists of multiple steps, and the corresponding prediction flow chart is illustrated in Fig. 1. Overall, the forecasting method is divided into two phases: a model-driven phase and a data-driven phase. In the model-driven phase, the shale multi-mineral petrophysical model, matrix parameters and corresponding apparent saturation are derived (Jin et al., 2019). Then, in the data-driven phase, the conventional logging curves, matrix parameters, and apparent saturation are used as inputs, whereas the output is the actual saturation (the actual saturation of the shale gas reservoir, different from the

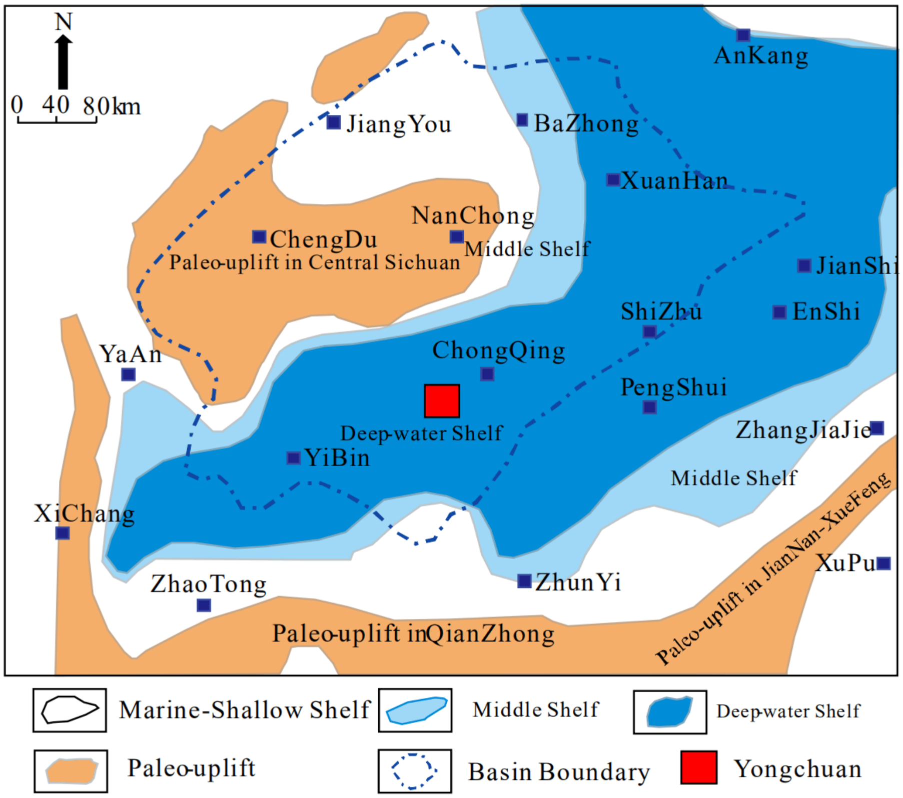


Fig. 2. Location of the research area.

apparent saturation), which is used to establish an intelligent fusion algorithm. The final actual saturation prediction model is obtained by combining the prediction results from a plurality of machine learning algorithms suitable for small samples. This procedure provides the saturation prediction algorithm with a theoretical basis, thereby improving the reliability of the model; furthermore, the proposed algorithm can significantly improve the prediction accuracy. The detailed model derivation and the corresponding different algorithms are explained in the appendix.

3. Research area and data

We selected the Longmaxi Formation-Wufeng Formation shale gas reservoir from the Yongchuan block of the Sichuan Basin for verification (Chen et al., 2018). The Yongchuan area is located in Yongchuan District, Chongqing. The Longmaxi Formation-Wufeng Formation shale gas reservoir is a typical ultra-deep, over-expanded, over-pressured shale gas reservoir with a depth of 3500-4500 m that was deposited in a deep-sea shelf-type environment. The reservoir is characterized by a broad distribution, considerable thickness, high total organic carbon (TOC) content, high vitrinite reflectance (Ro), brittleness, and good pore-fracture development. Evidently, the conditions under which the whole shale gas reservoir was formed and enriched were excellent. The main reservoir section is the first member of the Longmaxi Formation-Wufeng Formation and consists of shale rich in black carbon, silica, and carbonate with a thickness of 80-120 m. The TOC content of the Longmaxi section of the reservoir is in the range of 0-7.83%, the Gas Research Institute (GRI) total porosity range is 1.05-6.82%, the Ro range is 2-3%, and the clay content is 9.2-65.8%. A map depicting the location of the study area is shown in Fig. 2.

We applied the data from two wells, namely, wells YA and YB, in the Yongchuan block to test the model proposed in this paper. A total of 18 rock samples were drilled in well YA, and a total of 9 rock samples were drilled in well YB. Experiments were conducted to determine the TOC, GRI porosity, GRI saturation, and bulk density; in addition, X-ray diffraction was also performed to determine the clay content. Furthermore, we extracted the logging responses corresponding to the experimental depth interval. The GRI experimental methodology is as follows. 1) Extract a full-diameter core (approximately 300 g), weigh the core, use the core mercury injection technique to measure the total sample volume, and calculate the sample bulk density. 2) Crush the sample to a certain degree, then take the crushed sample and weigh it (approximately 100 g); subsequently, extract the toluene from the sample for 1 to 2 weeks, dry the sample at 110°C for another 1-2 weeks until the weight is stable, and then calculate the weight difference. 3) Using a helium medium to measure the volume of the particles after drying and calculate both the particle density and the total porosity. Based on this experiment, the water saturation can be calculated. The ranges of the parameters determined from the corresponding experiments are shown in Fig. 3.

We used well YA as the modelling well and well YB as the prediction well to analyse the proposed algorithm.

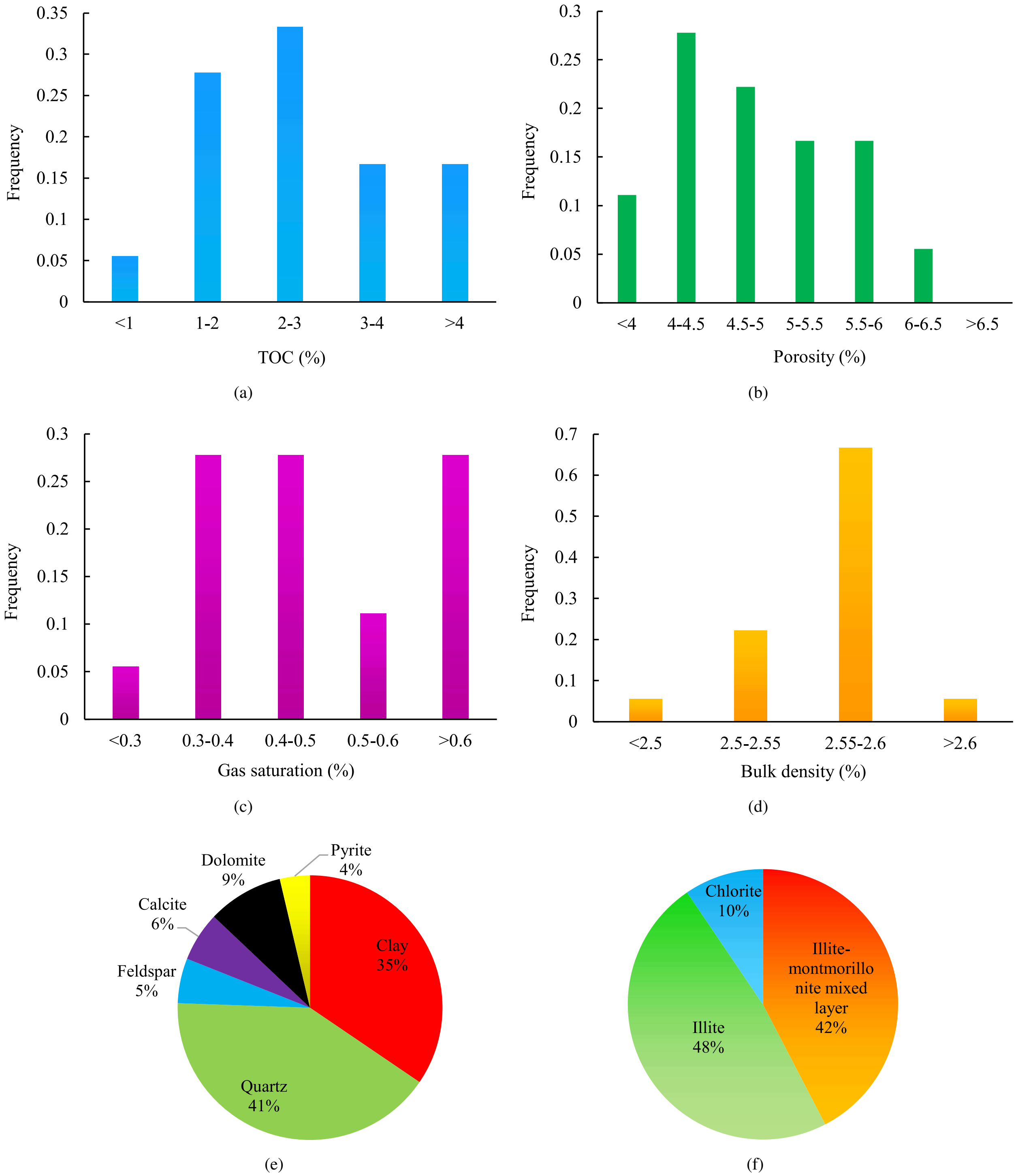


Fig. 3. Ranges of the parameters obtained from the core experiments. (a) Core TOC content range (0.74%-5.11%); (b) Core total porosity range (0%-6.351%); (c) Core gas saturation range (20.43%-72.31%); (d) Core bulk density range (2.66 g/cm^3-2.82 g/cm^3); (e) Distribution of each average mineral content in the shale samples involved in the experiment; (f) Distribution of each average clay content in the shale samples involved in the experiment.

Table 1. Theoretical response values for various rock components.

	Δt (us/ft)	ρ_b (g/cm^3)	N
Quartz	55.5	2.65	-0.02
Feldspar	51	2.68	-0.5
Calcite	46.5	2.71	0
Dolomite	41.5	2.87	0.03
Pyrite	39.2	4.997	-0.03
Natural gas	265	0.25	0.2
Water	189	1.05	1
Smectite	120	2.12	0.44
Illite	90	2.53	0.3
Chlorite	80	2.77	0.52

Since the exact petrophysical meaning of the entire prediction process is included in the model, the sample size required to participate in the modelling is greatly reduced. This reduced sample size is also consistent with the method proposed in this paper because it is unrealistic to perform various types of experiments on large quantities of rock samples and because the significance of the model for practical applications will diminish with a large sample size.

4. Modelling and experimental results

Before training the model and calculating the apparent saturation, we first assessed the relationships between the log responses and the GRI water saturation measured by the core experiment. Fig. 4 shows the relationships between the density log response, neutron log response, acoustic time difference (Δt) log response and core gas saturation/porosity/clay (because shale gas reservoirs contain only natural gas and water, 1 minus the water saturation gives the gas saturation).

Fig. 4 shows the relationships between the gas saturation, porosity, clay content, and porosity curve responses. Among them, the gas saturation, the density curve, the neutron porosity curve have relatively consistent relationships. In contrast, the relationship between the core porosity and porosity curve response is poor; this has been shown before because complex minerals affect the logging curve response, and thus, porosity logging cannot reflect the core porosity well. The clay content has the best relationship with the density curve, but this relationship is indirect because the clay content decreases as the organic matter content increases and because the organic matter density is much smaller than the density of other minerals, which causes the density curve response to decrease. Fig. 4 demonstrates that the correlation between Δt and the gas saturation is the worst with a coefficient of determination of 0.3856; hence, it is difficult to evaluate the gas saturation with Δt. In addition, the density log and neutron log responses are well correlated with the gas saturation because both density logging and neutron logging can reflect changes in the TOC content, which is indirectly related to gas saturation. However, since this relationship is indirect, the reliability of calculating the saturation directly using either density logging or neutron logging is insufficient. Therefore, these correlations are insufficient for evaluating gas saturation. Moreover, we did not choose to predict the reservoir gas saturation with the resistivity curve because the relationship between the resistivity curve and shale gas saturation is very complicated. Consequently, at present, no popular method exists for predicting shale gas saturation using resistivity curves. Furthermore, in addition to formation water (which is highly conductive) within the pore system, wet clay and pyrite minerals will also conduct electricity, further enhancing the difficult of applying the resistivity curve, particularly as the mineral contents and rock resistivity are not linear. Therefore, this paper does not consider using the resistivity characteristics of the reservoir.

We first calculated the apparent saturation, for which the theoretical responses of many components are needed. The responses for identifying the various mineral components, natural gas and formation water are shown in Table 1. The response values are determined by a common parameter table.

Table 1 shows the theoretical response values for the components commonly found in shale gas reservoirs. Among them, the response values of the different density logs are the most accurate. In addition, most of the values are accurate to two decimal places, although those of some components are accurate to three decimal places given the availability of density logging measurements. In contrast, the measurement accuracy of the neutron log and acoustic log response values is relatively limited, especially for the acoustic log responses; as a result, the acoustic responses of the clay minerals are obviously not accurate enough. This will affect the calculation of the saturation.

The density of organic matter changes with varying maturity. As the maturity increases, the density of organic matter also increases. Based on the findings of previous research, the density of organic matter was taken as 1.93 g/cm^3. In addition, the reference values of the organic matter neutron response and the acoustic wave time difference response given by Schlumberger are 0.65 pu and 120 us/ft, respectively. Using the above parameters, the matrix parameters are calculated, and the above results are used to calculate the apparent saturation. Fig. 5 depicts the relationships between the calculated matrix parameters and gas saturation, whereas Fig. 6 illustrates the

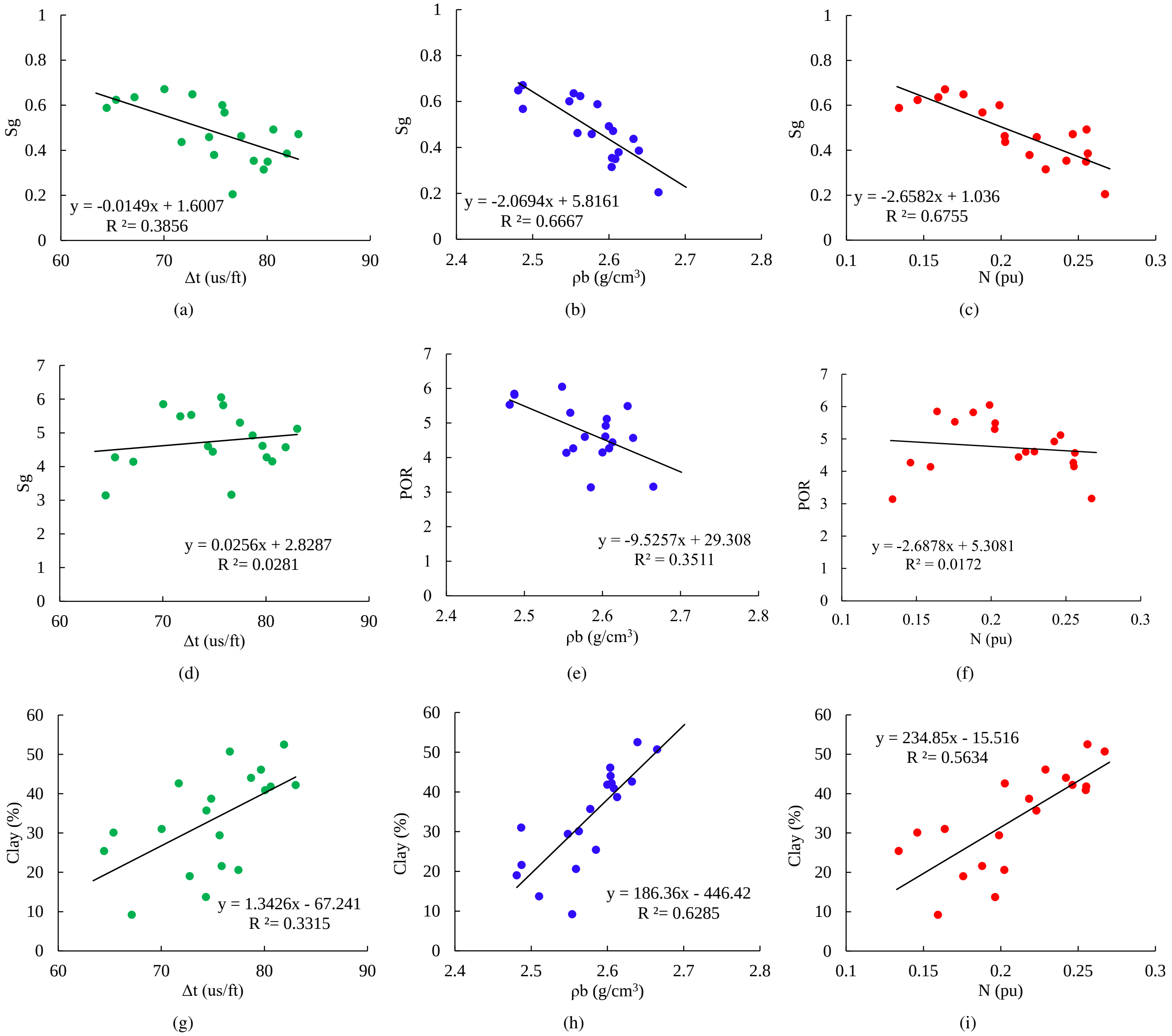


Fig. 4. Relationships between the logging responses and gas saturation. (a) Sg-Δt; (b) Sg-ρ_b; (c) Sg-N; (d) POR-Δt; (e) POR-ρ_b; (f) POR-N; (g) Clay-Δt; (h) Clay-ρ_b; (i) Clay-N.

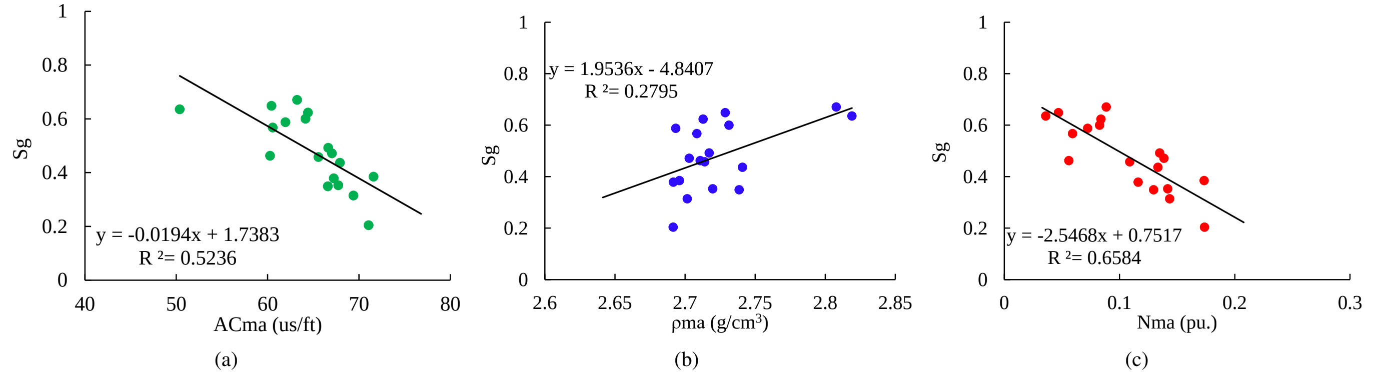
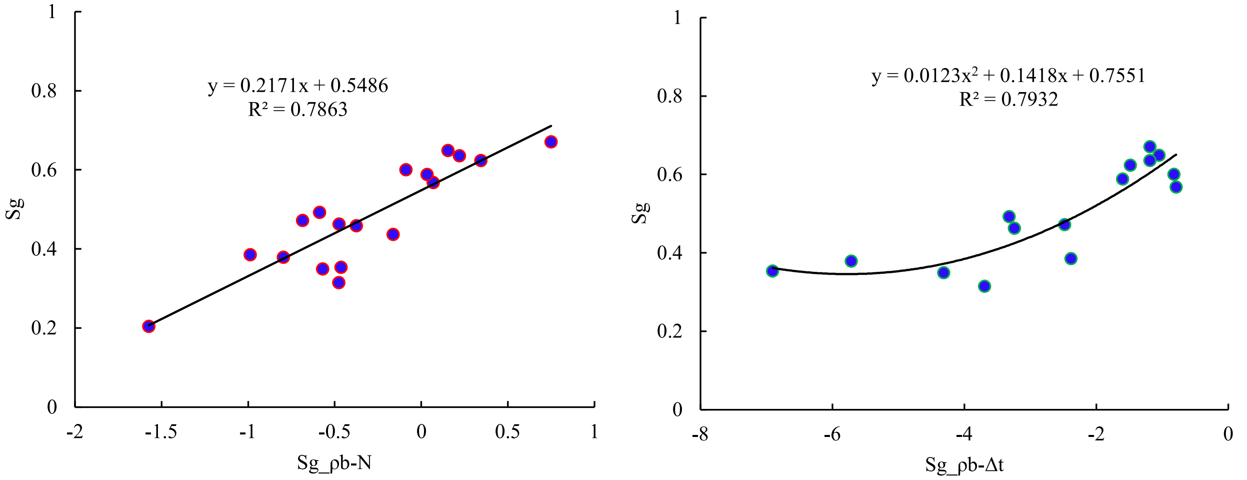


Fig. 5. Relationships between the matrix parameters and gas saturation.



(a) (b)

Fig. 6. Relationships between the apparent saturation and gas saturation.

relationships between the calculated apparent saturation and the gas saturation determined by core testing.

Fig. 5 demonstrates that the matrix parameters, namely, the acoustic wave time difference, neutron and density responses, are correlated with the gas saturation. Among them, the acoustic wave time difference and neutron responses have the best correlations with the gas saturation because the matrix minerals of the high-quality marine shale reservoir are composed primarily of biosilica; in addition, the corresponding clay content is reduced. The differences between the neutron and acoustic time difference responses of the clay content and biosilica are the largest, so these two matrix parameters (the acoustic wave time difference and the neutron responses) have good relationships with the gas saturation. This indirect relationship helps the saturation prediction and indicates that the measured logs are comprehensive responses to the entire rock system, while the matrix parameters exclude many other responses. The mineral information and measured logs also exhibit differences. The correlation between the density response and gas saturation is relatively poor because there is no information regarding the density of organic matter in the matrix. Only the density response of organic matter has a more direct relationship with the gas saturation. Nevertheless, the matrix density still prominently reflects the information regarding the clay content, so the matrix density parameter is still correlated with the gas saturation. Therefore, we added this curve to retain more information and improve the prediction effect.

Fig. 6 demonstrates that the apparent saturation calculated by the shale multi-mineral petrophysical model is well correlated with the measured saturation (with a coefficient of determination of 0.79 for both panels). Evidently, Sg_ρb-N and Sg have an approximately linear relationship, while S_ρb-Δt and Sg have a very nonlinear relationship. The effect of using a nonlinear fitting is better than that achieved using a linear fitting. Hence, if a nonlinear machine learning algorithm is used, the accuracy will definitely improve. On the basis of Fig. 6, adding the above information as inputs and employing a machine learning model for training, the prediction effect will also improve. Consequently, we used the back-propagation neural network (BPNN), extreme learning machine (ELM), random forest (RF), Adaboost, and bagging algorithms for simultaneous predictions. Several of these algorithms possess hyperparameters, and the determination of hyperparameters is very important; inappropriate hyperparameters can reduce the prediction accuracy of the model compared with the training accuracy. Therefore, we employed the training set-verification set method to determine the hyperparameters. To determine a certain hyperparameter combination, 20% of all samples are randomly extracted; then, the remaining 80% of the samples are used for training, and the 20% of extracted samples are predicted. The above three operations are repeated, and the sampling is re-randomized each time. The average of the three results is the prediction effect under the given hyperparameter, and the optimal hyperparameter is determined by traversing different combinations of hyperparameters.

BPNN and ELM are suitable for determining hyperparameters by distinguishing the validation set from the number of hidden layer neurons. The hyperparameters in the RF, Adaboost, and bagging algorithms that need to be determined consist mainly of the number of integrated base learners. The corresponding determination results are shown in Fig. 7. The formulas used to calculate the absolute error (AE), relative error (RE), and mean square error (MSE) are as follows:

$$\mathrm{AE} = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{Predicted}^{i} - \mathrm{Target}^{i}\right| \tag{1}$$

$$\mathrm{RE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\mathrm{Predicted}^{i} - \mathrm{Target}^{i}\right|}{\mathrm{Target}^{i}} \tag{2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Predicted}^{i} - \mathrm{Target}^{i}\right)^{2} \tag{3}$$

Fig. 7 shows the resulting hyperparameters determined by different algorithms. The RF, bagging and Adaboost algorithms based on tree structures are significantly less affected by hyperparameters than BPNN and ELM. This finding is attri-

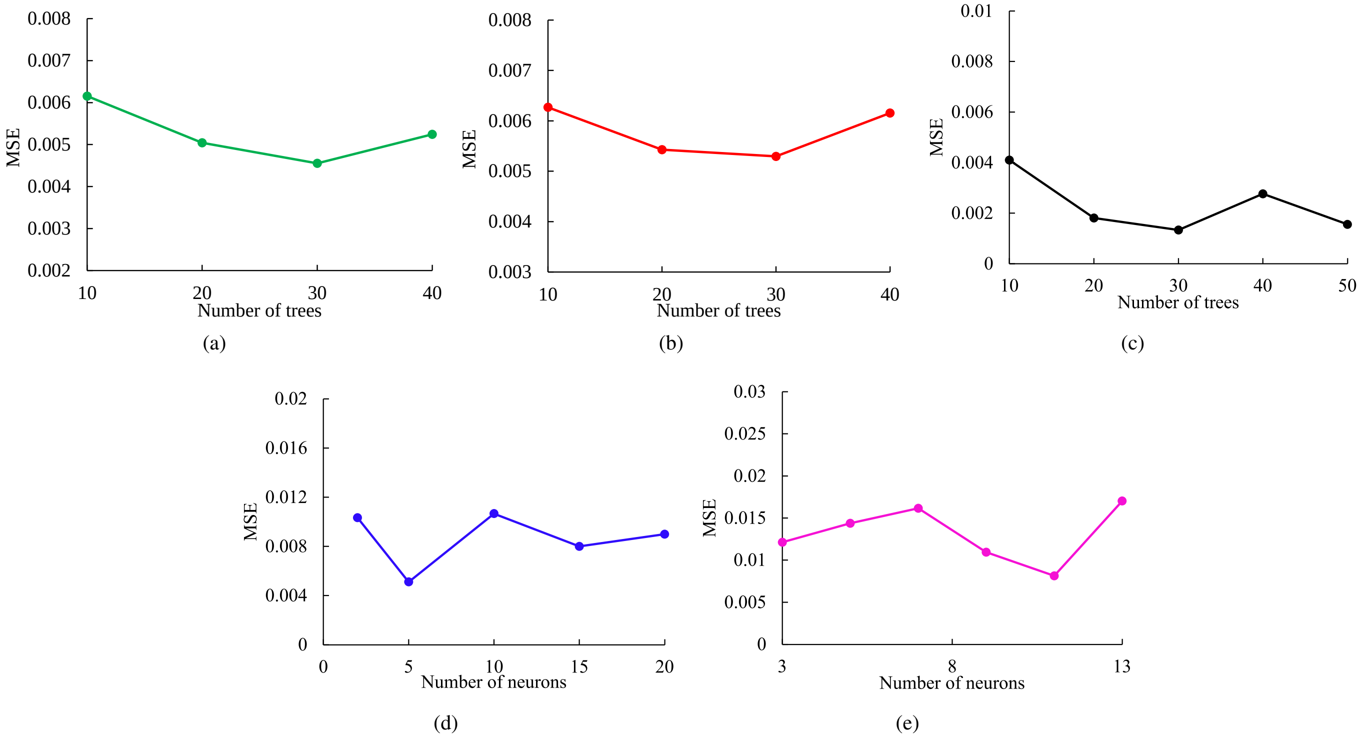


Fig. 7. Hyperparameter prediction results of the 5 algorithms. (a) RF; (b) Bagging; (c) Adaboost; (d) BPNN; (e) ELM.

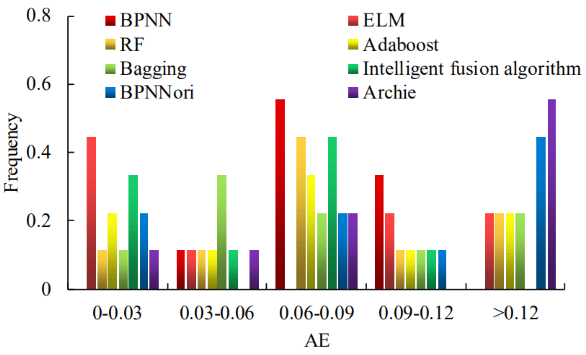


Fig. 8. Prediction error distribution for all 8 models.

buted to the fact that the ELM algorithm is more stable because it integrates multiple identical models, increasing its stability and reducing the risk of overfitting. The stability of the decision tree itself is higher than that of the neural network. Simultaneously, we can also see that the approximation ability of BPNN and ELM may be stronger than those of the other algorithms, especially that of the ELM, which achieves the best accuracy in the hyperparameter determination. Finally, the hyperparameter results are as follows: the optimal numbers of base learners for the RF, bagging and Adaboost algorithms are all 30; further, the optimal parameter for BPNN, the number of neurons in the hidden layer, is 5, and the number of neurons in the optimal hidden layer of ELM is 11.

We applied the model and the abovementioned hyperparameters to train the core data from well YA. After the training is completed, we predict the core data from well YB, which is a test well that does not participate in the modelling. If the proposed algorithm can perform better on the test well than the original algorithm, the attempt is deemed successful. The above five basic algorithms are compared with the intelligent fusion algorithm proposed in this paper. BPNN combines only three inputs, namely, density, neutron, and acoustic logging curves (model 7), and the saturation result is calculated by using the resistivity model (model 8). We use these 8 results to comprehensively analyse the proposed model and highlight its superiority. Among these algorithms, the number of hidden layer neurons in model 7 is set to 5, and model 8 uses the Archie formula to calculate the saturation. The corresponding formula is derived as follows:

$$S_g = 1 - S_w = 1 - \sqrt[n]{\frac{abR_w}{\varphi^m R_t}} \tag{4}$$

where a, b, m, and n are the Archie parameters. Considering the experimental results, a, b, m, and n are 1.3069, 1.0647, 1.446, and 1.549, respectively. The prediction error distribution among the eight models is shown in Fig. 8, and Fig. 9 shows the final core calculation results for the training well. The algorithm proposed in this paper has the highest accuracy, whereas the accuracy of using the data-driven method is relatively low but still better than that of the model-driven method. The intelligent fusion algorithm proposed in this paper and the prediction effect of model 7 are shown in Fig. 10, and the error statistics for each model are shown in Table 2. In Fig. 10 and Table 2, IFA refers to the intelligent fusion algorithm.

Fig. 9 confirms that the algorithm proposed in this paper can improve the prediction accuracy, and the prediction effect

Table 2. Theoretical response values for the various algorithms.

	BPNN	ELM	RF	Adaboost	bagging	IFA	BPNNori	Archie
AE	0.065	0.073	0.076	0.075	0.067	0.045	0.095	0.128
RE	0.125	0.124	0.13	0.142	0.124	0.094	0.178	0.398
MSE	0.007	0.008	0.007	0.007	0.006	0.004	0.012	0.025

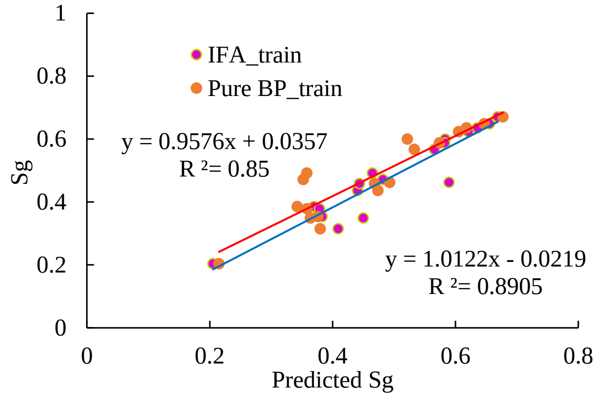
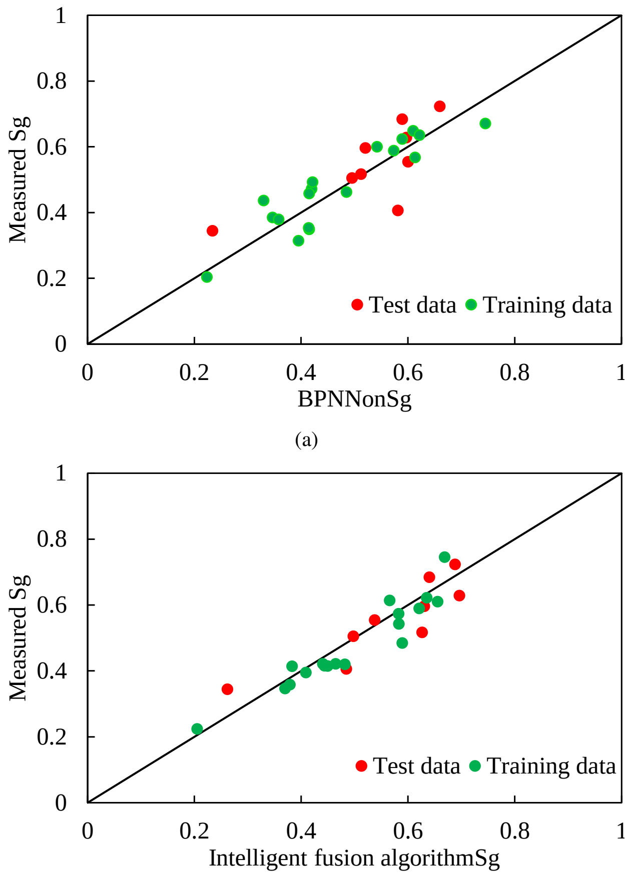


Fig. 9. Back-propagation results of the core data from the training well.



(a)

(b)

Fig. 10. Back-propagation results of the core data from the training well.

on the training data is significantly better than the effects achieved with the model-based and data-only methods. These findings show that the proposed intelligent fusion algorithm inherits both the reliability of the petrophysical model and the high precision of the data model.

In Fig. 10, BPNNoriSg refers to the result predicted by model 7, and the results predicted by the model proposed in this paper are denoted by "Intelligent fusion algorithm Sg". These results indicate that the prediction effect of the Archie formula based on resistivity is the worst: the AE of the prediction of most samples is greater than 0.12, and the RE and MSE are more than double those obtained for the gas saturation with the other prediction methods. Hence, the Archie model does not meet the requirements for accurate saturation calculations, and the gas saturation calculations using the Archie formula are not reliable for shale gas reservoirs. It is therefore necessary to study new conductivity models for the actual conditions of shale gas reservoirs. Using model 7, which is exclusively based on the data-driven model for predicting gas saturation, the prediction effect for the prediction well is insufficient. The calculated parameters cannot be applied to the calculation of reserves because the data-driven model cannot easily learn the true input-output intrinsic relationship in the case of poor processing, especially for problems with complex functional relationships and small sample sizes. In contrast to research on image and speech recognition, very few studies have been performed on targeted machine learning models for geoscience data, and thus, these models are difficult to control. The prediction system that combines a multi-mineral petrophysical model with the algorithm proposed in this paper is more accurate than the data-driven model. Consequently, the prediction results of the five benchmark algorithms reveal that the intelligent fusion algorithm provides a slightly better prediction effect than the BPNN algorithm. This finding is attributed to the fact that BPNN models are often somewhat sensitive and are not easily controlled in the case of small sample sizes, whereas integrated learning is more suitable for small sample prediction problems.

The intelligent fusion algorithm achieves the best results in predicting the gas saturation of the well, and the reliability of the prediction effect of the proposed method is much better than that of the other methods. Fig. 9 demonstrates that the model prediction results are highly accurate, and there is no deviation from the 45° line, indicating that the prediction is unbiased, which represents the role of the petrophysical model in the intelligent fusion algorithm. We believe that with the increased use of artificial intelligence and machine learning in professional settings, the integration of these methods can improve the prediction accuracy of reservoir parameters. However, it is important either to combine model-driven and

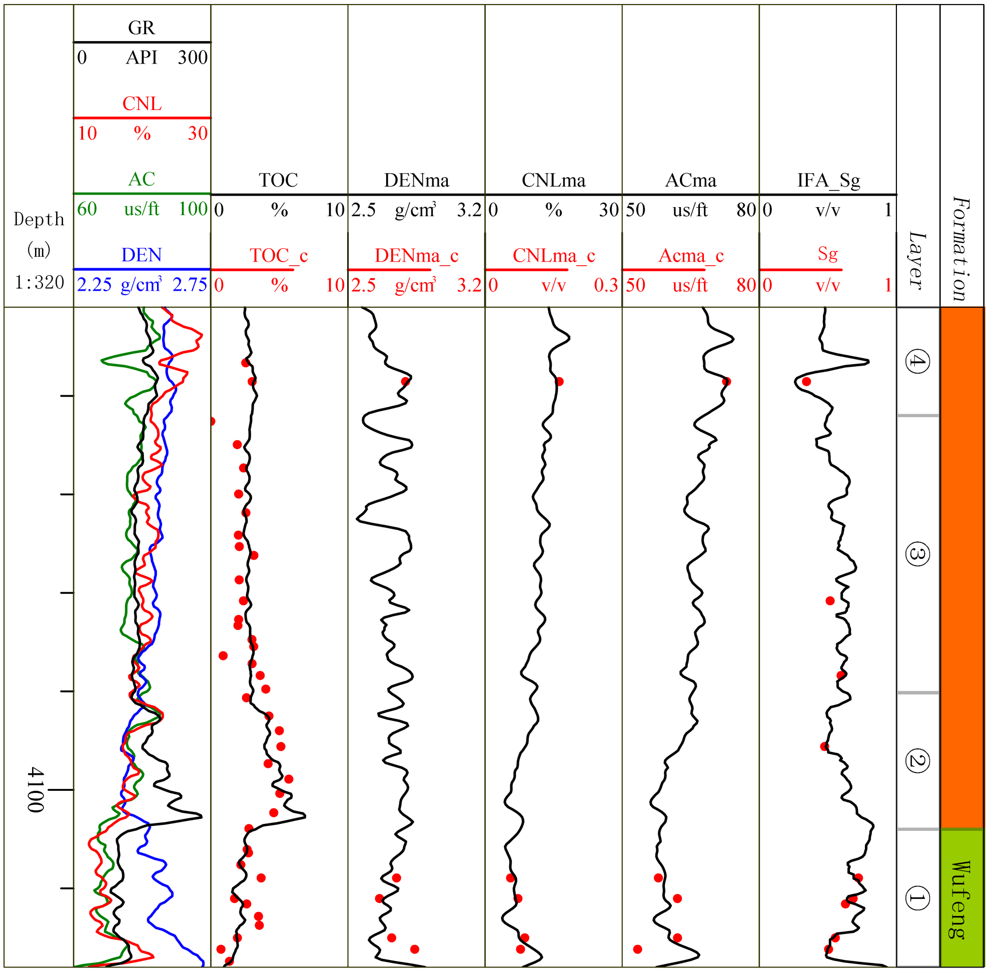


Fig. 11. Calculation results of gas saturation in the whole prediction well.

data-driven models or to propose a targeted machine learning model for specific problems. In summary, the model proposed in this paper considers the actual principles of petrophysics and fuses different algorithms to improve the accuracy of the model, which can provide significant help in calculating the gas saturation of shale gas reservoirs.

Fig. 11 characterizes the calculation results for the entire well. The penultimate track (on the far left) presents the evaluated gas saturation, where Sg is the core gas saturation value and IFA_Sg refers to the gas saturation calculated by the method proposed in this paper. First, the calculation accuracy of the proposed model is obviously very accurate for the core data points. Table 2 further indicates that the RE with the core logging calculation is less than 10%. Second, by observing the changes in the curve, even if well YB is a prediction well, the curve is very stable, and there are no severe fluctuations, yet there is no case where the curve predicts a fixed value. These results demonstrate that the proposed model is reliable and has strong generalization ability. This method can thus be applied to the calculation of gas saturation and can improve the gas saturation prediction accuracy. In the absence of a reliable shale gas resistivity saturation model, the intelligent fusion algorithm can be used for calculating the saturation of a shale gas reservoir.

5. Conclusions

In this paper, a high-precision method for calculating the gas saturation in organic shale pores with an intelligent fusion algorithm and a multi-mineral petrophysical model is proposed that combines two models of two different saturation calculation systems. The problem of inaccurate calculations of gas saturation in shale gas reservoirs is resolved, and the AE is reduced by more than 40% compared with the original prediction method. Based on the work presented in this article, we can summarize the following conclusions:

(1) When calculating the parameters of shale gas reservoirs, regardless of the model used, we recommend considering the complex mineral composition and organic matter of the shale gas reservoir. It is particularly difficult to employ simple statistical models to accurately calculate shale gas reservoir parameters.

(2) The shale gas reservoir saturation calculation method used at present is based mainly on the porosity curve. Moreover, the resistivity curve-based calculation saturation method is not ideal in shale gas reservoirs.

(3) The intelligent fusion algorithm proposed in this paper combines the prediction results of multiple machine learning models. Because the mathematical principles of different machine learning models are not consistent, the focus varies with the machine learning model. This fusion of multiple algorithms improves the gas saturation prediction effect.

(4) Combining the principles of petrophysics with the intelligent fusion algorithm and changing the inputs of the algorithm greatly improves the prediction effect of the model. For the problem of logging interpretation, the combination of model-driven and data-driven approaches can fully exploit the advantages of both methods.

Acknowledgement

This project was supported by the Yangtze University Excellent Doctoral and Master's Degree Thesis Cultivation Program; the National Natural Science Foundation of China (Nos. 41404084 and 41504094); the Open Fund of the Key Laboratory of Exploration Technologies for Oil and Gas Resources (Yangtze University); the Ministry of Education (Nos. K2015-06, K2016-09, K2017-01, and K2018-11); the Youth Innovation Team, Key Laboratory of Exploration Technologies for Oil and Gas Resources (Yangtze University), Ministry of Education (No. PI2018-03); the National Science and Technology Major Project (No. 2017ZX05032003-005); and the Yangtze University ESI Discipline Enhancement Program.

Conflict of interest

The authors declare no competing interest.

Open Access This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

Alfred, D., Vernik, L. A new petrophysical model for organic shales. Petrophysics 2013, 54(3): 240-247.

Cai, J., Wei, W., Hu, X., et al. Electrical conductivity models in saturated porous media: A review. Earth-Sci. Rev. 2017, 171: 419-433.

Cai, J., Xu, K., Zhu, Y., et al. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Appl. Energy 2020, 262: 114566.

Chai, J., Wang, M., Yang, B. Improved model of saturation based on Indonesian equation. Petroleum Geology and Engineering 2010, 24(2): 75-78. (in Chinese)

Chen, M., Dai, J., Liu, X., et al. Differences in the fluid characteristics between spontaneous imbibition and drainage in tight sandstone cores from nuclear magnetic resonance. Energy Fuels 2018, 32(10): 10333-10343.

Christou, V., Tsipouras, M.G., Giannakeas, N., et al. Hybrid extreme learning machine approach for heterogeneous neural networks. Neurocomputing 2018, 311: 397-412.

Danie, H.L., Gonzalo, M.M., Alberto, S. Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles. Neurocomputing 2011, 74(12-13): 2250-2264.

Dong, D., Zou, C., Dai, J., et al. Suggestions on the development strategy of shale gas in China. J. Nat. Gas Geosci. 2016, 1(6): 413-423.

Ge, X., Fan, Y., Cao, Y., et al. Investigation of organic related pores in unconventional reservoir and its quantitative evaluation. Energy Fuels 2016, 30(6): 4699-4709.

Guo, J., Xie, R., Xiao, L., et al. Nuclear magnetic resonance T1T2 spectra in heavy oil reservoirs. Energies 2019, 12(12): 2415.

He, Z., Hu, Z., Nie, H., et al. Characterization of shale gas enrichment in the Wufeng FormationLongmaxi Formation in the Sichuan Basin of China and evaluation of its geological constructiontransformation evolution sequence. J. Nat. Gas Geosci. 2017, 2(1): 1-10.

Ji, K., Guo, S., Hou, B. A logging calculation method for shale adsorbed gas content and its application. J. Pet. Sci. Eng. 2017, 150: 250-256.

Jiang, C., Zhang, S., Reyes, J. Black shale xenolith in a Jurassic-Cretaceous kimberlite and organic-rich Upper Ordovician shale on Bafan Island, Canada: A comparison of their organic matter. Mar. Pet. Geol. 2019, 103: 202-215.

Jiang, H., Zheng, W., Luo, L., et al. A two-stage minimax concave penalty based method in pruned Adaboost ensemble. Appl. Soft Comput. 2019, 83: 105674.

Jin, G., Xie, R., Lin, M., et al. Petrophysical parameter calculation based on NMR echo data in tight sandstone. IEEE Trans. Geosci. Remote Sens. 2019, 57(8): 5618-5625.

Josh, M., Piane, C.D., Esteban, L., et al. Advanced laboratory techniques characterizing solids, fluids and pores in shales. J. Pet. Sci. Eng. 2019, 180: 932-949.

Kang, Y., Shang, C., Zhou, H., et al. Mineralogical brittleness index as a function of weighting brittle mineralsfrom laboratory tests to case study. J. Nat. Gas Sci. Eng. 2020, 77: 103278.

Kim, G., Lee, H., Chen, Z., et al. Effect of reservoir characteristics on the productivity and production forecasting of the Montney shale gas in Canada. J. Pet. Sci. Eng. 2019, 182: 106276.

Lai, J., Pang, X., Xu, F., et al. Origin and formation mechanisms of low oil saturation reservoirs in Nanpu Sag, Bohai Bay Basin, China. Mar. Pet. Geol. 2019, 110: 317-334.

Li, J., Wu, Q., Jin, W., et al. Logging evaluation of free-gas saturation and volume content in Wufeng-Longmaxi organic-rich shales in the Upper Yangtze Platform, China. Mar. Pet. Geol. 2019, 100: 530-539.

Li, S., Cui, Z., Jiang, Z., et al. New method for prediction of shale gas content in continental shale formation using well logs. Appl. Geophys. 2016, 13: 393-405.

Li, W., Zou, C., Wang, H., et al. A model for calculating the formation resistivity factor in low and middle porosity sandstone formations considering the effect of pore geometry. J. Pet. Sci. Eng. 2017, 152: 193-203.

Li, X., Zhao, W., Zhou, C., et al. Dual-porosity saturation model of low-porosity and low-permeability clastic reservoirs. Pet. Explor. Dev. 2012, 39(1): 88-98.

Liu, S., Feng, M., Yan, W. Study on method of calculating water saturation of shale reservoir by non-electrical logging-taking the Jiaoshiba block of fuling shale gas field as an example. Science Technology and Engineering 2017, 17: 127-132. (in Chinese)

Malekimostaghim, E., Gholami, R., Rezaee, R., et al. A laboratory-based approach to determine Archie's cementation factor forshale reservoirs. J. Pet. Sci. Eng. 2019, 183: 106399.

Mashaba, V., Altermann, W. Calculation of water saturation in low resistivity gas reservoirs and pay-zones of the Cretaceous Grudja Formation, onshore Mozambique basin. Mar. Pet. Geol. 2015, 67: 249-261.

Mercadier, M., Lardy, J.P. Credit spread approximation and improvement using random forest regression. Eur. J. Oper. Res. 2019, 277(1): 351-365.

Morga, R., Kamińska, M. The chemical composition of graptolite periderm in the gas shales from the Baltic Basin of Poland. Int. J. Coal Geol. 2018, 199: 10-18.

Nazemi, M., Tavakoli, V., Rahimpour-Bonab, H., et al. The effect of carbonate reservoir heterogeneity on Archie's exponents (a and m), an example from Kangan and Dalan gas formations in the central Persian Gulf. J. Nat. Gas Sci. Eng. 2018, 59: 297-308.

Nazemi, M., Tavakoli, V., Sharifi-Yazdi, M., et al. The impact of micro-to macro-scale geological attributes on Archie's exponents, an example from PermianTriassic carbonate reservoirs of the central Persian Gulf. Mar. Pet. Geol. 2019, 102: 775-785.

Owusu, E.B., Tsegab, H., Sum, C.W., et al. Organic geochemical analyses of the Belata black shale, Peninsular Malaysia; implications on their shale gas potential. J. Nat. Gas Sci. Eng. 2019, 69: 102945.

Sang, Q., Zhang, S., Li Y., et al. Determination of organic and inorganic hydrocarbon saturations and effective porosities in shale using vacuum-imbibition method. Int. J. Coal Geol. 2018, 200: 123-134.

Shedid, S.A., Saad, M.A. Comparison and sensitivity analysis of water saturation models in shaly sandstone reservoirs using well logging data. J. Pet. Sci. Eng. 2017, 156: 536-545.

Tan, M., Mao, K., Song, X., et al. NMR petrophysical interpretation method of gas shale based on core NMR experiment. J. Pet. Sci. Eng. 2015, 136: 100-111.

Tariq, Z., Mahmoud, M., Al-Youssef, H., et al. Carbonate rocks resistivity determination using dual and triple porosity conductivity models. Petroleum 2020, 6(1): 35-42.

Tathed, P., Han, Y., Misra, S. Hydrocarbon saturation in upper Wolfcamp shale formation. Fuel 2018, 219: 375-388.

Wang, Y., Dong, D., Yang, H., et al. Quantitative characterization of reservoir space in the Lower Silurian Longmaxi Shale, southern Sichuan, China. Sci. China Earth Sci. 2014, 57: 313-322.

Wu, Y., Tahmasebi, P., Lin, C., et al. A comprehensive study on geometric, topological and fractal characterizations of pore systems in low-permeability reservoirs based on SEM, MICP, NMR, and X-ray CT experiments. Mar. Pet. Geol. 2019, 103: 12-28.

Zhang, C., Zhang, Z., Li, J., et al. Study on conductivity mechanism and saturation equation based on harbor effect. Journal of Oil and Gas Technology 2009, 31(6): 86-95. (in Chinese)

Zhao, P., Mao, Z. Improvement of Ramirez's petrophysical model for volume of shale gas reservoir. Acta Petrolei Sinica 2014, 35(3): 480-485. (in Chinese)

Zhao, P., Ma, H., Rasouli, V., et al. An improved model for estimating the TOC in shale formations. Mar. Pet. Geol. 2017, 83: 174-183.

Zhou, X., Zhang, C., Zhang, Z., et al. A saturation evaluation method in tight gas sandstones based on diagenetic facies. Mar. Pet. Geol. 2019, 107: 310-325.

Zhu, L., Zhang, C., Zhang, Z., et al. An improved theoretical nonelectric water saturation method for organic shale reservoirs. IEEE Access 2019a, 7: 51441-51456.

Zhu, L., Zhang, C., Zhang, Z., et al. Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. Appl. Soft Comput. 2019b, 83: 105596.

Zhu, L., Zhang, C., Zhang, C., et al. A new and reliable dual model- and data-driven TOC prediction concept: A TOC logging evaluation method using multiple overlapping methods integrated with semi-supervised deep learning. J. Pet. Sci. Eng., 2020, 188: 106944.

Zhu, X., Shan, S., Fu, D. Analysis and improvement on the conductivity model of low porosity and permeability shaly sand reservoirs. Progress in Geophysics 2016, 31(6): 2724-2728. (in Chinese)

Appendix A: Calculation of apparent saturation based on a shale multi-mineral petrophysical model

The calculation of reservoir parameters in a shale gas reservoir is more difficult than in a conventional reservoir given the diversity of pore types, complex pore structure, presence of organic matter and matrix mineral composition of the former, all of which impact the physical properties of shale rocks. In particular, variations in the skeletal mineral composition of shale and the differences in the organic matter properties of different shale gas blocks have led to significant discrepancies in the relationships between reservoir parameters and the logging responses of different shale gas blocks. Therefore, the reliability of the saturation calculation directly using a statistical model will be considerably reduced because the influences of other factors are not considered. In this paper, a theoretical model for calculating the apparent saturation is proposed for a shale multi-mineral petrophysical model to improve the theoretical nature of the saturation calculation and ultimately enhance the reliability of the model.

First, a multi-mineral petrophysical model of the shale reservoir is determined. To date, many scholars have proposed various shale petrophysical volume models to address different needs. For example, Wang et al. (2009) and Kang et al. (2020) suggested that a shale gas reservoir can be divided into three parts: a brittle mineral part, a clay mineral part, and an organic part. Among them, the brittle mineral part develops brittle mineral pores, the clay mineral part develops clay mineral pores, and the organic matter part develops organic pores. Similarly, Zhao et al. (2014) suggested that shale gas reservoirs can be divided into organic and inorganic fractions. Tan et al. (2015) proposed that a shale gas reservoir should be divided into five parts: the matrix part, the kerogen part, the adsorbed gas part, the free gas part and the bound water part. More recently, Li et al. (2016) suggested that shale rocks can be divided into solid kerogen and non-organic minerals.The common matrix components of shale gas reservoirs are quartz, feldspar, pyrite, calcite, dolomite, illite, chlorite. The density, hydrogen index and longitudinal wave time response vary for the above minerals, especially for clay minerals such as illite, chlorite, and illite-smectite mixed layers. Therefore, the matrix part needs to be divided among a plurality of mineral components. In addition, shale gas reservoirs contain organic matter (Ge et al., 2016; Zhao et al., 2017; Zhu et al., 2019b; Zhu et al., 2020), and the various physical responses of the organic matter interposed between the matrix and fluid must be distinguished. We did not separately determine the volume of bound water from that of adsorbed gas because the difference between adsorbed gas and natural gas is not substantial; in addition, a suitable method for measuring the volume of adsorbed gas by logging is currently not available. The response of bound water is also the same as that of movable water. The resulting shale multi-mineral petrophysical model is shown in Fig. A-1.

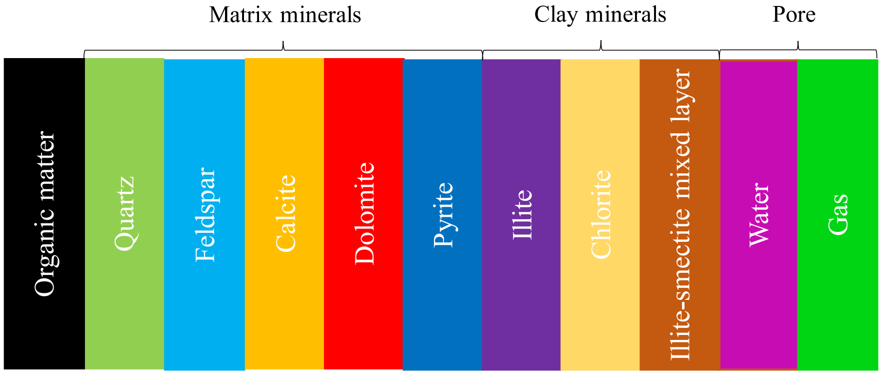


Fig. A-1. Multi-mineral petrophysical model (taking the shale rock in the article as an example).

It is worth mentioning that the parameter used to characterize the organic matter content is the TOC content, which has a conversion relationship with the organic matter content:

$$V_{OM} = k\left(\frac{\mathrm{TOC}}{\rho_{OM}}\right)\rho_b \tag{A-1}$$

where V_{OM} is the volume fraction of organic matter, TOC is the mass fraction of the total organic carbon content, ρ_{OM} is the density of organic matter, ρ_b is the density of rock, and k is the conversion coefficient of organic carbon related to the type of kerogen, etc., whose value can be taken as 1.2 or 1.25. Using the multi-mineral petrophysical model of Fig. A-1 combined with the log responses, we can list the theoretical formulas with the gas saturation term:

$$\rho_b = \rho_{ma}(1-\varphi-V_{OM}) + \rho_{OM}V_{OM} + \rho_w\varphi(1-S_h) + \rho_h\varphi S_h \tag{A-2}$$

$$N = N_{ma}(1-\varphi-V_{OM}) + N_{OM}V_{OM} + N_w\varphi(1-S_h) + N_h\varphi S_h \tag{A-3}$$

$$\Delta t=\Delta t_{ma}\left(1-\varphi-V_{OM}\right)+\Delta t_{OM}V_{OM}+\Delta t_{w}\varphi\left(1-S_h\right)+\Delta t_h\varphi S_h \tag{A-4}$$

where N is the rock neutron value, which is the neutron log response value; Δt is the rock acoustic wave time difference, which is the acoustic wave time difference log response value; ρ_{ma} is the skeletal density; N_{ma} is the matrix neutron value; Δt_{ma} is the difference between the matrix acoustic waves; φ is the porosity; ρ_{OM} is the density of organic matter; N_{OM} is the neutron value of organic matter; Δt_{OM} is the acoustic time difference of organic matter; ρ_w is the density value of the formation water; N_w is the neutron value of the formation water; Δt_w is the acoustic time difference of the formation water; S_h is the gas saturation; ρ_h is the density of natural gas; N_h is the neutron value of natural gas; and Δt_h is the acoustic time difference of natural gas. Among them, ρ_{ma}, N_{ma}, and Δt_{ma} vary greatly; this variation is related to the mineral composition and thus needs to be calculated in a targeted manner. Assuming that the shale rock matrix is composed of M mineral components, then the corresponding calculation formulas for the three parameters ρ_{ma}, N_{ma}, and Δt_{ma} are determined as follows:

$$\rho_{ma}=\rho_1V_1+\rho_2V_2+\cdots+\rho_MV_M \tag{A-5}$$

$$N_{ma}=N_1V_1+\rho_2V_2+\cdots+N_MV_M \tag{A-6}$$

$$\Delta t_{ma}=\Delta t_1V_1+\Delta t_2V_2+\cdots+\Delta t_MV_M \tag{A-7}$$

where ρ_i, i =1, 2, 3, $\cdots$ M, is the density value of the i-th mineral; N_i, i =1, 2, 3, $\cdots$ M, is the neutron value of the i-th mineral; Δt_i, i =1, 2, 3, $\cdots$ M, is the acoustic time difference of the i-th mineral; V_i, i =1, 2, 3, $\cdots$ M, is the volume fraction of the matrix occupied by the i-th mineral; and the cumulative volume of all M minerals is 1.

Here, we combine Eqs. (A-5), (A-6), and (A-7), perform a large number of equation transformations, eliminate the porosity parameter from the equations, and obtain a new saturation evaluation model as follows:

$$S_{h\_\rho_b-N}=\frac{\left(\rho_b-\rho_{ma}\left(1-V_{OM}\right)-\rho_{OM}V_{OM}\right)\left(N_w-N_{ma}\right)-\left(N-N_{ma}\left(1-V_{OM}\right)-N_{OM}V_{OM}\right)\left(\rho_w-\rho_{ma}\right)}{\left(N-N_{ma}\left(1-V_{OM}\right)-N_{OM}V_{OM}\right)\left(\rho_h-\rho_w\right)-\left(\rho_b-\rho_{ma}\left(1-V_{OM}\right)-\rho_{OM}V_{OM}\right)\left(N_h-N_w\right)} \tag{A-8}$$

$$S_{h\_\Delta t-\rho_b}=\frac{\left(\Delta t-\Delta t_{ma}\left(1-V_{OM}\right)-\Delta t_{OM}V_{OM}\right)\left(\rho_w-\rho_{ma}\right)-\left(\rho_b-\rho_{ma}\left(1-V_{OM}\right)-\rho_{OM}V_{OM}\right)\left(\Delta t_w-\Delta t_{ma}\right)}{\left(\rho_b-\rho_{ma}\left(1-V_{OM}\right)-\rho_{OM}V_{OM}\right)\left(\Delta t_h-\Delta t_w\right)-\left(\Delta t-\Delta t_{ma}\left(1-V_{OM}\right)-\Delta t_{OM}V_{OM}\right)\left(\rho_h-\rho_w\right)} \tag{A-9}$$

In this manner, two models for evaluating the gas saturation are obtained. Among them, $S_{h\_\rho_b-N}$ is the saturation calculated from the combination of density logging and neutron logging responses, whereas $S_{h\_\Delta t-\rho_b}$ is the saturation calculated from the combination of acoustic and density log responses. Eqs. (A-8)-(A-9) demonstrate that theoretically, if each parameter in the formula can be accurately determined, the gas saturation can be accurately calculated.

Although we have derived a model for calculating the gas saturation, the calculation formulas show that the accurate gas saturation values calculated by Eqs. (A-8)-(A-9) require the very accurate determination of a large number of parameters. This requirement makes the calculation very difficult, and numerous experiments are needed. Typically, we use only the theoretical values of each parameter in the formula, such as the response value of each matrix component and the response value of organic matter. This approach can greatly affect the prediction accuracy for shale gas reservoirs because the shale gas reservoir model contains an excessive number of parameters. Therefore, we assert that the gas saturation calculated by Eqs. (A-8)-(A-9) is only the apparent gas saturation. The calculated value has a certain theoretical significance and should exhibit a correlation with the measured saturation; this correlation is stronger than the correlations between the conventional logging curves and saturation. However, the calculated result still cannot represent the actual saturation, and thus, it is necessary to use the algorithm to approximate the measured saturation to improve the prediction accuracy.

Appendix B: Evaluation method based on intelligent multi-model fusion

Here, we introduce an evaluation method based on intelligent multi-model fusion. The prediction results of multiple models are considered together because this method requires a sample to perform various experiments when the model is determined. This type of sample is usually small, so it is necessary to use a variety of algorithms to improve the generalization ability of the total model and enhance the stability during prediction. We choose two neural network-based algorithms, namely, the error back-propagation neural network algorithm and extreme learning machine algorithm, and three integrated learning-based algorithms: the random forest algorithm, Adaboost regression algorithm and bagging regression algorithm. The final result will be obtained by averaging the above five methods. This process can effectively improve the predictive ability of the model, and there is no overfitting of the model. Below, we will introduce the training processes and principles of these algorithms.

Error back-propagation neural network

The error BPNN model is one of the most widely used algorithms in oil and gas exploration. BPNN, a function approximation algorithm, was originally proposed by Rumellhart in 1986 (Zhu et al., 2017). Structurally, BPNN consists of three parts: the input layer, the hidden layer, and the output layer. Each layer of the network is connected by a transfer function. The learning process of BPNN can be divided into two steps: signal forward propagation and error back propagation. In the first step, the signal is input by the input layer, and the calculated value of the neural network is output from the output layer by the weight between the neurons and the activation function in the neuron. In the second step, the error is back propagated, and the error between the output value of the output layer and the expected value, the comparison error, and the set learning precision are calculated. If the error is greater than the learning precision, the calculation error is used to obtain the partial derivative of the weight and the threshold in the neural network; furthermore, the weights between the neurons and the thresholds in the neurons are adjusted according to the gradient descent method. These two processes continue to run iteratively. If the number of iterations is greater than the set maximum number of iterations or if the error is less than the set learning accuracy, the BPNN training phase ceases.

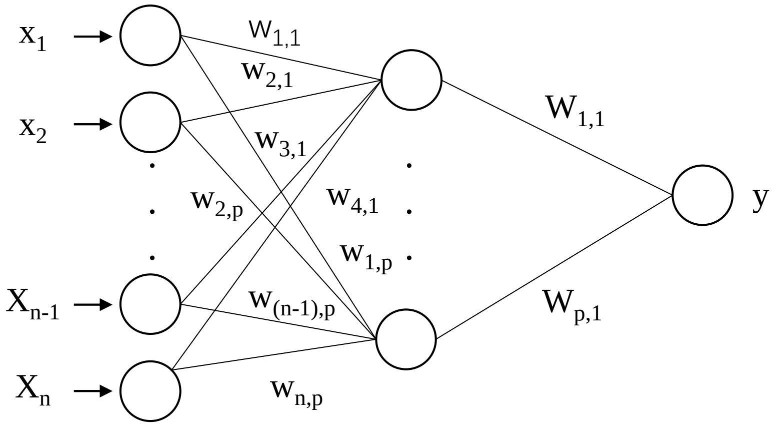


Fig. B-1. BPNN structure.

The BPNN network structure is depicted in Fig. B-1. In Fig. B-1, X_1, X_2, $\cdots$, X_h, $\cdots$, X_n are the input signals of the neural network, and $W_{h,i}$ represents the weights between the input-layer neurons and the hidden layer neurons. In addition, α_1, α_2, $\cdots$, α_h, $\cdots$, α_p are the inputs to the neurons in the hidden layer, W_h represents the weights of the neurons in the hidden layer and the neurons in the output layer, and y is the output value of the neural network.

Extreme learning machine

The ELM algorithm, which was initially proposed by Guangbin Huang in 2004 (Christou et al., 2018), is characterized by randomly or artificially assigned weights of the hidden layer nodes and very fast training and prediction phases. The hidden layer weights do not need to be updated, and the random weights and learning process calculate only the output weight. A schematic diagram of the ELM structure is shown in Fig. B-2.

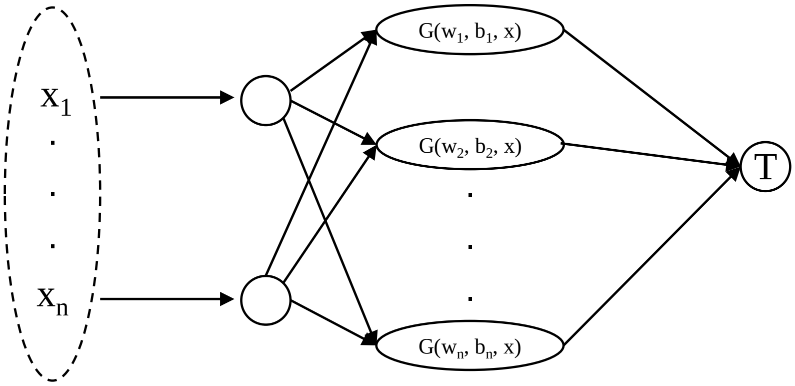


Fig. B-2. ELM structure.

For a training sample (x, t), the output function of the single hidden layer forward neural network with L hidden layer neurons is determined as follows:

$$f_L(x)=\sum_{i=1}^{L}\beta_i G(a_i,b_i,x) \tag{B-1}$$

where a_i and b_i are hidden layer node parameters, β_i represents the weight between the i-th hidden layer and the network output, and $G(a_i,a_i,x)$ denotes the hidden layer node output of the i-th hidden layer corresponding to the sample x. For the additive hidden layer node, $G(a_i,a_i,x)$ is expressed as follows:

$$G(a_i,b_i,x)=g(a_i x+b_i) \tag{B-2}$$

where g is the activation function and $a_i x$ represents the inner product of the weight vector a_i and the sample x. Assuming that the connection function of the hidden layer uses a Gaussian function, $G(a_i,b_i,x)$ is calculated as follows:

$$G(a_i,b_i,x)=g(b_i\|x-a_i\|) \tag{B-3}$$

where a_i and b_i (b_i>0) represent the centre of the i-th Gaussian function node and the influence factor, respectively. Then, the entire prediction equation can be expressed as follows:

$$H\beta=T \tag{B-4}$$

where H is the hidden layer output matrix and T is the training target.

The error function of ELM uses the L_2 mean square error function: after introducing the L_2 regularization term, the error function is:

$$\min_{\beta\in R_{L\times m}}\frac{1}{2}\|\beta\|^2+\frac{C}{2}\|H\beta-T\|^2 \tag{B-5}$$

C is a regularization coefficient, and its solution is expressed as follows:

$$\beta^*=\left(H^T H+\frac{1}{C}\right)^{-1}H^T H \tag{B-6}$$

Random forest algorithm

The RF algorithm, which was proposed by Breiman and Cutler in 2001, is a predictor containing multiple decision trees (Mercadier and Lardy, 2019; Cai et al., 2020). If the decision tree is a classification tree, the RF algorithm can be used for classification; if the decision tree is a regression tree, the RF algorithm can be used for regression. RF is an extended variant of bagging. RF uses a decision tree to build the bagging integration for the base learner and further introduces random attribute selection into the training process of the decision tree (Fig. B-3).

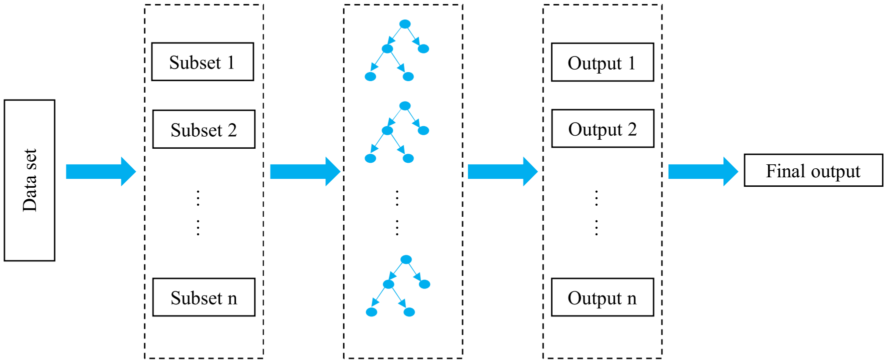


Fig. B-3. RF structure.

RFs are built in a random manner. There are many decision trees in the forest, and there is no correlation between each pair of decision trees in the RF. After obtaining the forest, when a new sample is input, each decision tree in the forest is predicted separately, and the result is a synthesis of all decision tree results.

In the process of establishing each decision tree, there are two points that should be considered: sampling and complete splitting. The former is the process of two random samplings, where the RF algorithm samples the rows and columns of the input data. To sample the rows, there is a method to return the data back to the data set. Specifically, in the sample set obtained by sampling, there may be duplicate samples. Assuming that there are N input samples, then there are also N sampled samples. As a result, not all of the input samples of each tree are samples during training, making it relatively easy to lead to overfitting. Then, column sampling is performed. From M features, m features ($m \ll M$) are selected. This process is followed by a decision tree that completely splits the sampled data. The final result is obtained from the predicted average given by each decision tree. The principle of RF is very simple, and the algorithm is easy to implement and works well on most problems, which is why this algorithm is applied herein.

Adaboost algorithm

The Adaboost algorithm is an integrated learning algorithm proposed by Freund and Schapire (Jiang et al., 2019b). The corresponding structure is shown in Fig. B-4.

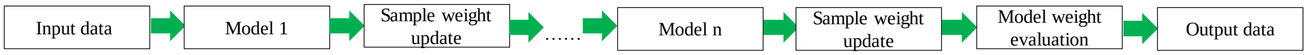


Fig. B-4. Adaboost structure.

The key property of the Adaboost algorithm is to transform a weak learning algorithm into a strong learning algorithm. This feature provides an effective new concept and design for learning algorithms when it is very difficult to construct a strong learner. Unlike the RF algorithm and the bagging algorithm, the Adaboost algorithm applies all samples when training each learner. The model structure of Adaboost is shown in Fig. B-4. Adaboost trains the training model in serial mode. After each training step, each sample is assigned a weight, where the weights of samples with lower prediction accuracy are increased. Thus, in the next training, the samples with lower prediction accuracy can be targeted. After several models are trained, the prediction errors of the samples outside the bag are assigned different weights for different learners. Fig. B-4 illustrates the Adaboost structure during training. The initial weight of the i-th sample, the weight of each base learner, and the weight of the different samples when updating the next iteration are calculated as follows:

Set the initial weight of the ith sample as $D_1(i)$, which is defined as follows:

$$D_1(i) = \frac{1}{N} \tag{B-7}$$

Under D_1, the base predictor $h_1(x)$ is trained, and the Bayesian regularization error function is used to calculate each sample error ε_i and the average error ε_t. The weight of the current base predictor is calculated using the sample error ε_i and the average error ε_t obtained for each training, and the weights of different samples at the next iteration are updated.

$$\begin{cases} W_t = \frac{1}{2}\ln\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right) \\ D_{t+1}(i) = \dfrac{D_t(i)\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{-\varepsilon_i}}{\sum\limits_{i=1}^{n}\left(D_t(i)\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{-\varepsilon_i}\right)} \end{cases} \tag{B-8}$$

where W_t is the weight of the t-th classifier and $D_{t+1}(i)$ is the weight of the t+1-th BPNN sample.

Bagging regression algorithm

The bagging algorithm is the most famous representation of the parallel integrated learning method (Danie et al., 2011). The RF algorithm mentioned above is adjusted on the basis of bagging. Given a data set containing m samples, we first randomly take a sample into the sample set and return the sample back to the original data set. Thus, the sample may be selected the next time the data set is sampled. After m random sampling operations, we obtain a sample set with m samples. Some samples in the initial training set appear in the sample set multiple times, whereas others never appear.

Thus, we can sample T samples with m training samples. Then, we train a base learner based on each sample set and combine these base learners. This is the basic flow of bagging. When combining the predicted outputs, bagging uses a simple averaging method for regression tasks.

The main difference between bagging and boosting is that bagging focuses predominantly on reducing the variance, while boosting focuses mainly on reducing the deviation. The focus of each algorithm is different. In theory, the larger the model difference, the better the fusion effect. When the actual prediction is made, the prediction results of the above five models will be averaged to obtain the final prediction result and improve the stability of the prediction.